Les gens se méfient de l'intelligence artificielle et, d'une certaine manière, cela a du sens. Avec le désir de créer les modèles d'IA les plus performants, de nombreuses organisations ont donné la priorité à la complexité plutôt qu'aux concepts d'explicabilité et de confiance.

Comme le monde devient de plus en plus dépendant des algorithmes pour prendre un large éventail de décisions, les dirigeants de technologies et d'entreprises seront chargés d'expliquer comment un modèle a choisi son résultat.

La transparence est une condition essentielle pour générer la confiance et l'adoption de l'IA. Parallèlement, la conformité réglementaire et la sécurité des modèles exigent également des entreprises qu'elles conçoivent un certain niveau d'interprétabilité dans les modèles. En outre, une entreprise doit évaluer les données qu'elle utilise pour s'assurer que les systèmes n'apprennent pas et ne renforcent pas des biais inconscients. Les organisations peuvent avoir besoin d'augmenter les données existantes pour créer un échantillon représentatif et tenir compte des changements dans les lois, les normes sociétales et le langage.

Des algorithmes d'IA complexes permettent aux organisations de tirer des enseignements de données qui étaient auparavant inaccessibles. Cependant, la nature de ces systèmes, qui sont des boîtes noires, signifie qu'il n'est pas facile pour les utilisateurs de comprendre la logique qui sous-tend la décision. Ce n’est pas tant un enjeu pour des applications de reconnaissance d’images qui classent ou ne classent pas bien des objets (ce qui ne nécessitent pas d’explication). Par exemple, l'expert comptable à l’origine du logiciel Tiime qui classe automatiquement grâce à son application Tiime Receipt les justificatifs de dépenses afférentes à une facture automatiquement dans le bon compte n’a de sujet ‘explicatibilité » car son appli IA réussit à classer la dépense ou ne réussit pas (auquel cas, elle le dit), mais n’a pas besoin de se justifier pour avoir la confiance de son utilisateur. En revanche, si ce cabinet venait à développer une application qui formulerait des recommandations en matière de gestion d’entreprise, l’utilisateur (le dirigeant) serait intéressé que l’application lui motive ses recommandations sur base de fait et/ou d’analyse spécifiques.

Même les spécialistes des données qui ont créé le modèle peuvent avoir du mal à expliquer pourquoi leur algorithme a pris une décision particulière. Une façon d'obtenir une meilleure

transparence des modèles est d'adopter une famille spécifique de modèles considérés comme explicables. Ces familles comprennent par exemple les modèles linéaires, les arbres de décision, les ensembles de règles, les ensembles de décisions, les modèles additifs généralisés et les méthodes de raisonnement basées sur des cas.

Des modèles plus simples peuvent souvent fournir un bon équilibre entre l'explicabilité et la performance (par exemple, précision, meilleure compréhension). Pourtant, de nombreuses entreprises n'aiment pas cette approche car elles estiment qu'un modèle plus complexe donne des résultats supérieurs. En raison de ces difficultés, les chercheurs et les fournisseurs de technologies travaillent sur les moyens d'aider les entreprises à expliquer les modèles.

Il existe un large éventail d'universitaires, de chercheurs et de fournisseurs de technologies qui travaillent sur les principes et les outils permettant de créer une IA fiable. Preet Gandhi de Nvidia a décrit deux classes de techniques d'explicabilité dans son billet de blog KDNuggets - ante-hoc et post-hoc. "Les techniques ante-hoc consistent à intégrer l'explicabilité dans un modèle dès le départ. Les techniques post-hoc permettent aux modèles d'être formés normalement, l'explicabilité n'étant incorporée qu'au moment des tests". Au cours des six derniers mois, j'ai interviewé un mélange de ces personnes pour en savoir plus sur les progrès réalisés dans le domaine de l'éthique de l'IA.

L'un des leaders dans ce domaine est Aleksandra (Saška) Mojsilović, qui est une scientifique, . Elle est également membre de l'IEEE et de l'IBM Academy of Technology. Saška est l'auteur de plus de 100 publications et détient 16 brevets. Mojsilović s'attaque de front aux questions de la confiance dans l'IA dans le cadre de son travail chez IBM.

Selon Mme Mojsilovic, l'IA de confiance doit être équitable, facile à comprendre et responsable. En outre, l'utilisateur doit avoir la certitude que le modèle est sûr et n'a pas été altéré. L'équité algorithmique est un sujet brûlant dans les cercles de l'IA. Elle a également noté que "lorsqu'il s'agit d'expliquer les décisions prises par les algorithmes, il n'y a pas une seule approche qui fonctionne le mieux. Il y a plusieurs façons d'expliquer. Le choix approprié dépend de la personnalité du consommateur et des exigences de la filière de l'IA".

Par exemple, il y a une grande différence entre les types d'explications souhaitées par les personnes qui produisent des modèles d'apprentissage machine tels que les scientifiques et les chercheurs spécialisés dans les données, et les décideurs commerciaux et les utilisateurs finaux qui consomment les informations produites par ces modèles. Un producteur de modèles

peut vouloir des explications qui l'aident à améliorer les modèles. Un chef d'entreprise veut comprendre des questions telles que les sources de données et la fiabilité de ces résultats. Une personne qui a reçu une décision basée sur un modèle veut savoir quels sont les facteurs qui ont contribué à la réponse.

À cette fin, Mojsilović et l'équipe d'IBM ont lancé l 'AI Explicability 360 en août 2019. AI Explainability 360 est une boîte à outils open-source extensible qui utilise diverses techniques pour expliquer et interpréter le modèle de prise de décision AI. AI Explainability 360 comprend 8 algorithmes d'explicabilité de pointe provenant de la recherche IBM, ainsi que des algorithmes supplémentaires provenant de la communauté de recherche plus large. Il existe également des tutoriels spécifiques à l'industrie. La boîte à outils AI Explainability 360 propose différents algorithmes d'explicabilité pour de multiples publics. Par exemple, dans les services financiers, un agent de crédit peut vouloir mieux comprendre comment le système d'IA a abouti à la recommandation d'approuver ou de refuser un prêt. Un client d'une banque, en revanche, peut avoir besoin d'une explication différente sur les raisons du refus ou de l'approbation d'un prêt.

Des outils tels que AI Explicability 360 aident les entreprises à comprendre les facteurs qui ont influencé la décision. Par exemple, si le modèle refuse une nouvelle demande de carte de crédit, la banque devrait pouvoir dire au consommateur que des facteurs tels qu'un paiement récemment manqué, un historique de crédit court et un faible pointage de crédit ont été les éléments essentiels de la décision.

Un autre sujet brûlant de l'IA est l'élimination des biais qui peuvent se matérialiser de nombreuses manières et à partir de nombreuses sources différentes. La boîte à outils AI Fairness 360 d'IBM, lancée en 2018, est une boîte à outils logicielle à code source libre qui peut aider à détecter et à éliminer les biais dans les modèles d'apprentissage automatique. La boîte à outils fournit aux développeurs des algorithmes à utiliser pour vérifier l'absence de biais indésirables. Par exemple, dans le domaine des services financiers, une banque voudrait savoir comment les prévisions de son modèle affecteront différents groupes de personnes (comme l'origine ethnique, le sexe ou le statut de handicap).

Des outils tels que Fairness 360 aideraient la banque à supprimer du modèle les paramètres susceptibles de provoquer un biais, tels que le code postal et l'appartenance ethnique. Les entreprises qui construisent des modèles d'IA doivent concevoir des modèles explicables et équitables dès le départ et les tester tout au long du cycle de vie du modèle pour s'assurer que des résultats équitables et fiables sont produits.

Mais qu'en est-il du cloud ?

De plus en plus, les entreprises utilisent des outils résidant dans le nuage pour construire des modèles d'IA. Les entreprises doivent également faire preuve de confiance et de transparence dans leurs services de cloud computing et les modèles d'IA produits par ces services. Tracy Frey, directrice de la stratégie pour le Cloud AI chez Google, est un autre cadre dirigeant qui mène la charge pour fournir une meilleure AI en améliorant l'interprétabilité de l'AI. Frey s'engage à ce que l'IA dans les nuages de Google soit responsable, réfléchie et collaborative, tout en continuant à faire progresser l'intelligence artificielle et l'apprentissage machine. Avant de rejoindre Google, elle a travaillé dans plusieurs start-ups technologiques en phase de démarrage, où elle a occupé plusieurs fonctions, notamment la gestion des produits, les relations avec les développeurs, le marketing des produits, le développement commercial et la stratégie.

En novembre 2019, la société a annoncé les explications de Google Cloud AI. Les explications quantifient la contribution de chaque facteur de données à la production d'un modèle d'apprentissage machine. Ces résumés aident les entreprises à comprendre pourquoi le modèle a pris les décisions qu'il a prises. Les organisations peuvent utiliser ces informations pour améliorer encore leurs modèles ou partager des informations utiles avec les consommateurs du modèle. Comme l'a noté Frey dans un article de blog, "Bien sûr, toute méthode d'explication a ses limites. D'une part, les explications d'IA reflètent les modèles que le modèle a trouvés dans les données, mais elles ne révèlent aucune relation fondamentale dans votre échantillon de données, votre population ou votre application".

La société a également introduit ce qu'elle appelle les Model Cards, en commençant par des cartes pour la détection des visages et des objets dans son offre d'API Cloud Vision. Les "Model Cards" sont de courts documents accompagnant des modèles d'apprentissage de machine formés qui fournissent des informations pratiques sur les performances et les limites des modèles". L'objectif de ces cartes est d'aider les développeurs à prendre de meilleures décisions sur les modèles à utiliser et sur la manière de les déployer de manière responsable.

Les chefs d'entreprise et les consommateurs doivent pouvoir faire confiance aux résultats des modèles d'intelligence artificielle avancés. Idéalement, une entreprise souhaite intégrer l'explicabilité dans la conception initiale d'un modèle. Cependant, il est tout aussi important d'expliquer les résultats des modèles existants pour garantir l'équité, vérifier la cohérence et permettre d'améliorer les cycles de recyclage.

2020 : La montée en puissance de la gouvernance de l'IA

Alors que nous nous dirigeons vers l'adoption de modèles d'IA plus avancés avec un apprentissage approfondi, il est impératif que nous disposions également d'un ensemble d'outils qui peuvent nous aider à mettre en évidence les problèmes, non seulement au moment où nous créons le modèle mais aussi au fur et à mesure de son évolution. Tout comme la gouvernance des données est devenue une pratique, nous verrons les organisations créer un rôle et adopter des outils pour la gouvernance de l'IA, ne serait-ce que pour atténuer le risque réglementaire potentiel. Bien qu'il faille du temps pour créer une IA plus explicable, il est prometteur de voir tant de technologues de tous les domaines travailler à la résolution des problèmes liés à l'IA.